Research Paper | Engineering | E-ISSN No : 2454-9916 | Volume : 3 | Issue : 12 | Dec 2017

ANALYTICS DRIVEN INFLUENCING METHODS FOR ELECTION STRATEGIES

Rutweek Sawant

B.E. Student of Computer Engineering, Atharva College of Engineering, Mumbai University, Mumbai, MH, India.

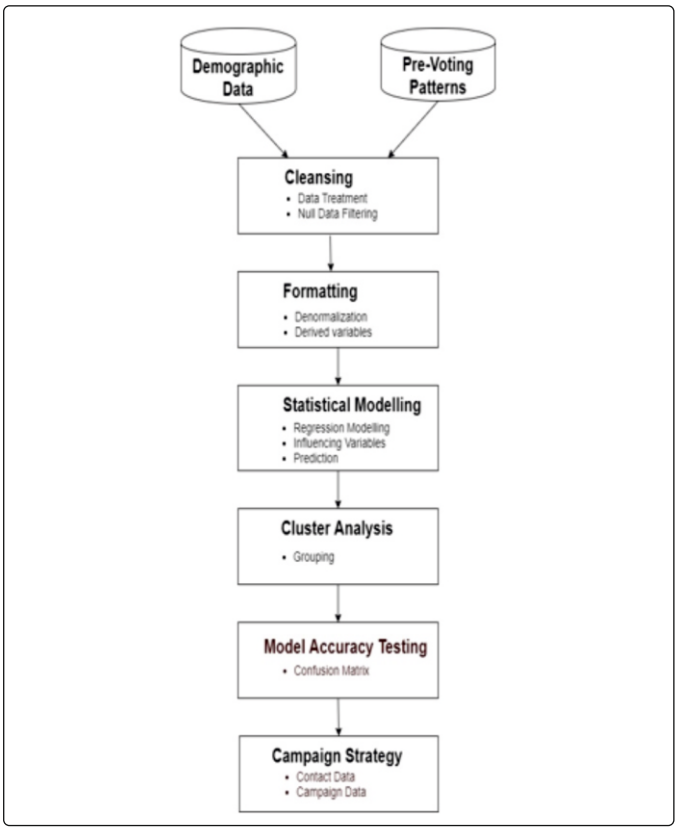
ABSTRACT

In today's VUCA (volatility, uncertainty, complexity and ambiguity) world, to become winner, it is essential to keep up with the advancements and to inundate them in our work. Technology, not only helps us in bettering the current scenario but also to understand our future. It is not only applicable in health sciences, business, environment, etc., but also in culture, education and even politics. In this project, we will see how analytics helps us to understand the voters, analyze their patterns and figure out the factors that can influence their votes ranging from communities, religion, education, employment etc., which finally would help to secure maximum percentile of votes by a candidate/political party. Thus, this model uses analytics in vote prediction with statistical modelling consisting of regression analysis such as logistic regression and random forest approach, finding influential variables, prediction via hypothesis, accuracy testing using confusion matrix and a way to cluster using probability analysis.

KEYWORDS: logistic regression; random forest; influential variables; prediction; hypothesis; confusion matrix; cluster.

I. INTRODUCTION:

It is always better to know the basket well, before putting your eggs in it.' Similarly, in the case of a political election, before investing in the heavy campaigning, it is better to know if doing that is likely to yield the maximum possible votes or not. For a political party or contestant, one of the strategies could be heavy campaigning on a selective population synonymously a medium level contact program; but the real question is which strategy is suitable for which group of population because in a diverse country like India. Voting patterns and results are highly unpredictable and hence a lot of energy gets focused on generic and divisive strategies for vote share. Analytics provides a very structured and scientific way of addressing this challenge. It can give the political parties the ability to have focused strategies and influence outcomes. Analytics provides a solution by bifurcating the voters into two groups on the basis of the probability of them voting the desired candidate/political party. Once this bifurcation is made, it will provide a detailed insight into which strategies to put into place to address the voters who are unlikely to vote in order to secure their votes. Also, they can generate strategies for those who have high probability of voting and ensure that they definitely vote.



II. PROPOSED SYSTEM:

I propose a framework that will classify the voter data available into 2 main groups that is High Probable Data – Voters likely to vote the Political party "A" or Low Probable Data – Voters unlikely to vote the political party "A" according to our Hypothesis equation. As shown in the figure, detailed analysis of the desired output is based upon the following phases.

- Voter Data Assimilation
- Data Cleansing
- Data Formatting
- Statistical Modelling
- Cluster Analysis
- Model Accuracy Testing
- Campaign Strategy

Voter Data Assimilation: Consists of raw data required by our model. Voter Data consists of 2 crucial information sets.

- Demographic Data
- Pre-Voting Patterns

Demographic Data: This data is of an individual voter pertaining to various aspects regarding his/her age, caste, community, income group, residing location, employment status, education status, influence susceptibility etc. These aspects from the core classification attributes to our prediction model. For example, Which political party, the people of income below 2 lakhs per annum would vote? What is the probability of 'X' community voting political party 'A'.

Pre-Voting Patterns: As demographic data forms an integral part for prediction so are the pre-voting patterns. They mainly provide the insight of whom the individual has voted during past elections. The more data available of an individual of his/her past voting the more accurate the output result will be. Pre-voting patterns are termed as highly confidential and government do not provide this data openly hence they have to be generated using various surveys.

Data Cleansing: The Data gleaned from various data tables contains various inconsistencies such as different formats, incomplete/null data values. Therefore, our primary goal is to clean it from these inconsistencies. For this, variables with significant missing values as well as variables with low variance are removed as they do not provide any predictive power. Sometimes it may be found that low variance variables are missing in a row for an individual entry but high variance variables are present, hence assumptions are filled in null entities.

Data Formatting: It is the process of converting the existing data in a single standard form. Its main goal is to improve the read performance by creating derived variables. It consists of 2 steps

- Denormalization
- Derived Variables

Denormalization: It is the process of concatenating various tables such as tables

of age, education status, caste, religion into a single table as pure direct information cannot be found in a single survey.

Derived Variables: Derived variables are obtained by grouping 2 or more columns to generate meaningful data. For example. If a table consists 2 distinct columns such as age group and income. We can generate a derived variable column by grouping age groups and income such as income for age group below 20 years.

Statistical Modelling: It is mathematical modeling procedure used for prediction and consists of following procedures.

- Regression Analysis
- Prediction

Regression Analysis: It is a statistical modelling process which is used to find the influencing variables. The efficiency of predictor model is directly proportional to the predictors, independent variables, dependent variables and the most important amongst all, the shape of the regression line. Following are the types of regression analysis that can be used

- Logistic Regression
- Random Forest

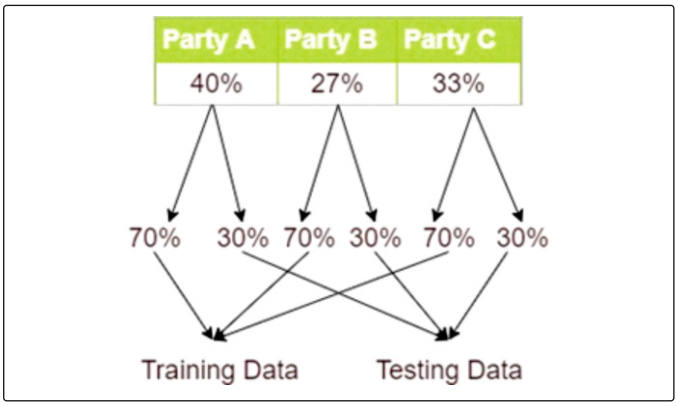
Logistic regression: It is a regression analysis technique in which on the basis of predictor variables the output of dependent variables is predicted.

Random Forest: It is an ensemble method. It uses averaging to improve the predictive accuracy. It processes by fitting decision tree classifier on various subsamples of dataset.

Influencing Variables: These variables help us to understand meaningful relations between independent variables and dependent variables.

Prediction: It is a statistical process in which we generate a hypothesis, it means we generate a testing condition in form of an equation. For example, political party 'A' winning the elections or voters of certain age group and graduate level education status voting political party "A". After generating this equation, we train some part of the data against this conditional equation and then test the remaining data for accuracy. Given below is a hypothesis test condition and its prediction.

Age-Group	Education-Status	Employement-Status	Sector	Hypothesis
30-50	Graduate	Employed	Private	High
50-55	Post-Graduate	Employed	Private	Low



In the above given diagram total votes are divide between 3 political party and the individual party votes are further bifurcated into 2 samples.

Training Data: It is used to train an algorithm and understand the characteristics of model. The better quantity of training data, the better algorithm performs. Usually 70% of data is taken as sample.

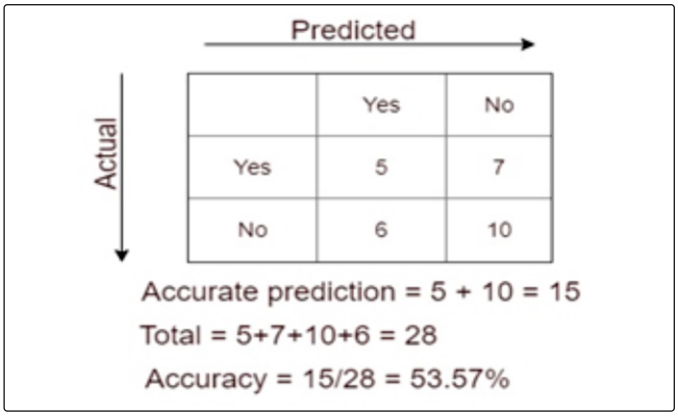
Testing Data: After training a data set, its is usually tested on testing data set to calculate the accuracy as the data in the testing set contains the known prediction output making it easy to test the model's accuracy. Usually 30% data is used for testing as a sample size.

Clustering: It means grouping similar objects in one class. Over here voters are clustered into various groups on the basis of their probability of fulfilling them the assigned hypothesis. To perform clustering, we then use grid based method. In this we map hypothesis condition vs probability and segregate the voters into low probability and high probability buckets.

Voter	P(Hypothesis)	Probability	
V1	High	0.56	✓
V2	High	0.84	✓
V3	Low	0.49	✗
V4	Low	0.12	✗

Model Accuracy Testing: This is used to test the accuracy of predictions. It is done with the help of confusion matrix as it maps actual results against predicted results.

	Actual Voted Party	Predicted Voted Party
Voter 1	A	A
Voter 2	A	B
Voter 3	B	A
Voter 4	B	B



Campaign strategy: This is the final step. In this step the clustered data is grouped and the groups are segregated into contact and campaign data. For example, Low and Medium Probability data is termed as Campaign data and the high probability data is termed as Contact data. These groups are of utmost important as the required output are elaborated by them. For example, questions such as, what is the probability that community 'X' will vote political party 'A'? are answered.

Campaign Data: This data consists of voters/voter groups that are less likely to vote the desired political party. In technical terms these voters have less probability of fulfilling our generated hypothesis equation. Hence these voters are of crucial importance because the real goal of this model is to find the voters that are unlikely to vote and address these voters, thereby securing their vote. Following are different ways to address campaign data voters.

- Seminars
- Rallies
- Awareness programs
- Sports and community events
- Televised events and speeches

Contact Data: This data consists of voters/voter groups that are most likely to vote our desired political party. In technical terms these voters have the highest probability of fulfilling our generated hypothesis equation. So unlike campaign data voters these kinds of voters need to be address periodically and the main objective should be to make sure that these voters do not change their votes. Following are the ways to address contact data voters periodically.

- SMS
- Election manifesto
- Emails
- Pamphlets

III. CONCLUSION:

This model gives a scientific solution for selecting strategies those can address specific needs of homogenous voter groups such as communities, age groups, income groups, common interest etc. It eliminates risk of selecting inappropriate campaigning exercises for voters. This model differentiates voters between campaign/contact data. Finally, it ensures winning voters trust successfully.

REFERENCES:

1. "Optimizing the Data Warehouse Design by Hierarchical Denormalizing", Morteza Zaker, Somnuk Phon-Amnuaisuk, Su-Cheng Haw, Proceedings of the 8th WSEAS International Conference on APPLIED COMPUTER SCIENCE (ACS'08)
2. "Linear Regression Analysis", Astrid Schneider, Dipl.Math.,Gerhard Hommel, Prof. Dr. rer. nat.,and Maria Blettner, Prof. Dr. rer. nat.,Dtsch Arztebl Int. 2010 Nov; 107(44): 776–782.Published online 2010 Nov 5. doi: 10.3238/arztebl.2010.0776
3. "Correlation and Simple Linear Regression" , Kelly H. Zou, PhD, Kemal Tuncali, MD, Stuart G. Silverman MD, Published online 10.1148/radiol. 2273011499 Radiology 2003, 227:617–628
4. "A Study of Hierarchical Clustering Algorithm", Yogita Rani and Dr. Harish Rohil, International Journal of Information and Computation Technology. ISSN 0974-2239 Volume 3, Number 11 (2013), pp. 1225-1232
5. "Performance Analysis of Hierarchical Clustering Algorithm", K. Ranjini, Dr. N. Rajalingam, Int. J. Advanced Networking and Applications Volume: 03, Issue: 01, Pages: 1006-1011 (2011)
6. "Implementing and Improvisation of K-means Clustering", Unnati R. Raval, Chaita Jani, IJCSMC, Vol. 4, Issue. 11, November 2015
7. "An Optimized Approach for k-means Clustering", Sadhana Tiwari, Tanu Solanki, International Journal of Computer Applications (0975 – 8887) 9th International ICST Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine-2013)
8. "Analysis of a Random Forests Model", Gerard Biau, Journal of Machine Learning Research 13 (2012) 1063-1095
9. "Random Foret: A Review", Eesha Goel, Er. Abhilasha, Internationals Journal of Advanced Research in Computer Science and Software Engineering Volume 7, Issue 1, January 2017